UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/055,586	01/22/2002	Ziv Bar-Yossef	ARC920010068US1	6945

23334 7590 03/02/2005

FLEIT, KAIN, GIBBONS, GUTMAN, BONGINI
& BIANCO P.L.
ONE BOCA COMMERCE CENTER
551 NORTHWEST 77TH STREET, SUITE 111
BOCA RATON, FL 33487

EXAMINER	
WASSUM, LUKE S	

ART UNIT	PAPER NUMBER
2167	

DATE MAILED: 03/02/2005

Please find below and/or attached an Office communication concerning this application or proceeding.

Office Action Summary	Application No.	Applicant(s)	
	10/055,586	BAR-YOSSEF ET AL.	
	Examiner	Art Unit	
	Luke S. Wassum	2167	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If the period for reply specified above is less than thirty (30) days, a reply within the statutory minimum of thirty (30) days will be considered timely.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *24 November 2004*.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1,2,4-14 and 16-20* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☐ Claim(s) *1,2,4-14 and 16-20* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *28 February 2002* is/are: a)☐ accepted or b)☒ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☐ Information Disclosure Statement(s) (PTO-1449 or PTO/SB/08) Paper No(s)/Mail Date _____.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____.
5) ☐ Notice of Informal Patent Application (PTO-152)
6) ☐ Other: _____.

DETAILED ACTION

Response to Amendment

1. The Applicants' amendment, filed 24 November 2004, has been received, entered into the record, and considered.

2. As a result of the amendment, claims 1, 4, 7-9, 11, 13, 16, 19 and 20 have been amended, and claims 3 and 15 have been canceled. Claims 1, 2, 4-14 and 16-20 are now pending in the application.

The Invention

3. The claimed invention is a system and method for cleaning a set of hypertext documents in order to minimize violations of a Hypertext Information Retrieval rule set.

Drawings

4. The drawings are objected to because in view of the Applicants' remarks with regard to the pending rejection to claims 4-6 and 16-18, the decision branch in Figure 6 from decision box 612 that is currently labeled "NO" should be labeled "No/Cannot be determined" (or some analogous correction to be proposed by the Applicants). This is because according to the specification and the Applicants' remarks, this path is taken in the case where *v* is determined not to be a pagelet, and also when it is not possible to determine whether or not *v* is a pagelet.

Corrected drawing sheets in compliance with 37 CFR 1.121(d) are required in reply to the Office action to avoid abandonment of the application. Any amended replacement drawing sheet should include all of the figures appearing on the immediate prior version of the sheet, even if only

one figure is being amended. The figure or figure number of an amended drawing should not be labeled as "amended." If a drawing figure is to be canceled, the appropriate figure must be removed from the replacement sheet, and where necessary, the remaining figures must be renumbered and appropriate changes made to the brief description of the several views of the drawings for consistency. Additional replacement sheets may be necessary to show the renumbering of the remaining figures. Each drawing sheet submitted after the filing date of an application must be labeled in the top margin as either "Replacement Sheet" or "New Sheet" pursuant to 37 CFR 1.121(d). If the changes are not accepted by the examiner, the applicant will be notified and informed of any required corrective action in the next Office action. The objection to the drawings will not be held in abeyance.

Claim Rejections - 35 USC § 112

5. In view of the Applicants' remarks regarding the pending rejection of claims 4-6 and 16-18 under 35 U.S.C. § 112, the examiner withdraws the rejections, under the condition that the above drawing objection be resolved.

6. The following is a quotation of the second paragraph of 35 U.S.C. 112:

> The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention.

7. Claims 1, 2, 4-7, 9-14 and 16-19 are rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

8. Regarding claims 1, 9, 11 and 13, each of these claims contains the limitation that templates comprise collections of pagelets wherein all pagelets are identical or almost identical. This limitation renders the claims indefinite, since it would be impossible for one of ordinary skill in the art to determine what precisely qualifies as 'almost identical', and therefore the exact meets and bounds of patent protection conveyed by the claims would be unclear.

9. Claims 2, 4-7, 10, 12, 14 and 16-19, fully incorporating the deficiencies of their respective parent claims, are likewise rejected.

Claim Rejections - 35 USC § 103

10. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

11. The factual inquiries set forth in *Graham* v. *John Deere Co*, 383 U.S. 1, 148 USPQ 459 (1966), that are applied for establishing a background for determining obviousness under 35 U.S.C. 103(a) are summarized as follows:

1. Determining the scope and contents of the prior art.
2. Ascertaining the differences between the prior art and the claims at issue.
3. Resolving the level of ordinary skill in the pertinent art.
4. Considering objective evidence present in the application indicating obviousness or nonobviousness.

12. This application currently names joint inventors. In considering patentability of the claims under 35 U.S.C. 103(a), the examiner presumes that the subject matter of the various claims was commonly owned at the time any inventions covered therein were made absent any evidence to the contrary. Applicant is advised of the obligation under 37 CFR 1.56 to point out the inventor and invention dates of each claim that was not commonly owned at the time a later invention was made in order for the examiner to consider the applicability of 35 U.S.C. 103(c) and potential 35 U.S.C. 102(e), (f) or (g) prior art under 35 U.S.C. 103(a).

13. Claims 1, 2, 7-9, 11, 13, 14, 19 and 20 are rejected under 35 U.S.C. 103(a) as being unpatentable over **Broder et al.** ("Syntactic Clustering of the Web") in view of **Huang** ("A Survey on Web Information Retrieval Technologies").

14. Regarding claim 1, **Broder et al.** teaches a method as claimed, comprising the step of cleaning, by operations of a computer system, a set of text documents to minimize violations of a predetermined set of Hypertext Information Retrieval rules (see section 1 Introduction, beginning on page 2; see also section 5.1 Common Shingles, beginning on page 7, particularly the disclosure on page 8 that common shingles either have no effect on the overall resemblance of the documents or they have the effect of creating a false resemblance between two basically dissimilar documents, and so common shingles are ignored), wherein the cleaning step comprises the steps of:

a) decomposing each page of the set of text documents into one or more pagelets (see disclosure that documents are decomposed into shingles, analogous to the claimed pagelets, in section 2 Defining Similarity of Documents, beginning on page 3);

b) identifying all pagelets belonging to templates (see disclosure that common shingles were nearly all mechanically generated, including shared header or footer information on a large number of automatically generated pages, i.e. forms, analogous to the claimed templates, in section 5.1 Common Shingles, beginning on page 7); and

c) eliminating the template pagelets from a data set (see section 5.1 Common Shingles, beginning on page 7, particularly the disclosure on page 8 that common shingles either have no effect on the overall resemblance of the documents or they have the effect of creating a false resemblance between two basically dissimilar documents, and so common shingles are ignored).

Broder et al. does not explicitly teach a method wherein a template comprises a collection of pagelets satisfying the requirements that all pagelets are identical or almost identical, and every two pages owning pagelets are reachable one from the other via other pages also owning pagelets in the collection.

Huang, however, teaches a method for eliminating duplicate web pages wherein a template comprises a collection of pagelets satisfying the requirements that all pagelets are identical or almost identical, and every two pages owning pagelets are reachable one from the other via other pages also owning pagelets in the collection (see section 3.5.2 Duplicate Elimination, pages 17-19, and particularly the disclosure at the top of page 18 that the formal definition for a similar collection includes requirements that that the pages are similar, and also that the links are similar, meaning that each of the pages should have at least one parent in their corresponding collection that are also similar pages; see also disclosure of the use of hyperlinks in categorizing documents, section 4.2.3

Enhanced Categorization Using Hyperlinks, specifically Radius-Two Specialization, which uses co-citation as a criterion for classification, on page 24).

It would have been obvious to one of ordinary skill in the art at the time of the invention to define templates with regard to similar pagelets and also similar links, since this would help to ensure that the templates being scrutinized are in fact similar, given the fact that one of the major difficulties in detecting replicated collections is that many replicas may not be strictly identical to each other (page 17).

15. Regarding claim 8, **Broder et al.** teaches a method as claimed, comprising the step of cleaning, by operations of a computer system, a set of text documents to minimize violations of a predetermined set of Hypertext Information Retrieval rules (see section 1 Introduction, beginning on page 2; see also section 5.1 Common Shingles, beginning on page 7, particularly the disclosure on page 8 that common shingles either have no effect on the overall resemblance of the documents or they have the effect of creating a false resemblance between two basically dissimilar documents, and so common shingles are ignored), wherein the cleaning step comprises the steps of:

a) decomposing each page of the set of text documents into one or more pagelets (see disclosure that documents are decomposed into shingles, analogous to the claimed pagelets, in section 2 Defining Similarity of Documents, beginning on page 3);

b) identifying all pagelets belonging to templates (see disclosure that common shingles were nearly all mechanically generated, including shared header or footer information on a

large number of automatically generated pages, i.e. forms, analogous to the claimed templates, in section 5.1 Common Shingles, beginning on page 7); and

c) eliminating the template pagelets from a data set (see section 5.1 Common Shingles, beginning on page 7, particularly the disclosure on page 8 that common shingles either have no effect on the overall resemblance of the documents or they have the effect of creating a false resemblance between two basically dissimilar documents, and so common shingles are ignored), wherein the identifying step comprises:

i) calculating a shingle value for each page and for each pagelet in the document set (see disclosure that documents are decomposed into shingles, analogous to the claimed pagelets, in section 2 Defining Similarity of Documents, beginning on page 3; see also disclosure of the fingerprint function, refered to as the shingle value, page 6); and

ii) sorting the pagelets by their shingle values into clusters (see disclosure in section 4 Algorithms, whereby similar documents are clustered, page 6).

Broder et al. does not explicitly teach a method wherein a template comprises a collection of pagelets satisfying the requirements that all pagelets are identical or almost identical, and every two pages owning pagelets are reachable one from the other via other pages also owning pagelets in the collection.

Huang, however, teaches a method for eliminating duplicate web pages wherein a template comprises a collection of pagelets satisfying the requirements that all pagelets are identical or almost identical, and every two pages owning pagelets are reachable one from the other via other pages also

owning pagelets in the collection (see section 3.5.2 Duplicate Elimination, pages 17-19, and particularly the disclosure at the top of page 18 that the formal definition for a similar collection includes requirements that that the pages are similar, and also that the links are similar, meaning that each of the pages should have at least one parent in their corresponding collection that are also similar pages; see also disclosure of the use of hyperlinks in categorizing documents, section 4.2.3 Enhanced Categorization Using Hyperlinks, specifically Radius-Two Specialization, which uses co-citation as a criterion for classification, on page 24).

It would have been obvious to one of ordinary skill in the art at the time of the invention to define templates with regard to similar pagelets and also similar links, since this would help to ensure that the templates being scrutinized are in fact similar, given the fact that one of the major difficulties in detecting replicated collections is that many replicas may not be strictly identical to each other (page 17).

16. Regarding claim 9, **Broder et al.** teaches a system as claimed, comprising:

a) a user interface (see disclosure that the system can be used for filtering the results of Web searches, said web searches requiring a user interface, Abstract);

b) a user interface/event manager communicatively coupled to the user interface (see disclosure that the system can be used for filtering the results of Web searches, said web searches requiring an event handler to respond to user requests, Abstract);

c) a generic data gathering device (see disclosure that the system can be applied to a group of documents found by the AltaVista spider, section 1 Introduction, beginning on page 2);

d) a generic information retrieval application, communicatively coupled to the user interface/event manager (see disclosure that the system can be applied to a group of documents found by the AltaVista spider, section 1 Introduction, beginning on page 2); and

e) a data cleaning application communicatively coupled to the generic data gathering application and to the generic information retrieval application for

i) decomposing each page of a set of text documents into one or more pagelets (see disclosure that documents are decomposed into shingles, analogous to the claimed pagelets, in section 2 Defining Similarity of Documents, beginning on page 3);

ii) identifying all pagelets belonging to templates (see disclosure that common shingles were nearly all mechanically generated, including shared header or footer information on a large number of automatically generated pages, i.e. forms, analogous to the claimed templates, in section 5.1 Common Shingles, beginning on page 7); and

iii) eliminating the template pagelets from a data set (see section 5.1 Common Shingles, beginning on page 7, particularly the disclosure on page 8 that common shingles either have no effect on the overall resemblance of the documents or they have the effect of creating a false resemblance between two basically dissimilar documents, and so common shingles are ignored).

Broder et al. does not explicitly teach a method wherein a template comprises a collection of pagelets satisfying the requirements that all pagelets are identical or almost identical, and every two pages owning pagelets are reachable one from the other via other pages also owning pagelets in the collection.

Huang, however, teaches a method for eliminating duplicate web pages wherein a template comprises a collection of pagelets satisfying the requirements that all pagelets are identical or almost identical, and every two pages owning pagelets are reachable one from the other via other pages also owning pagelets in the collection (see section 3.5.2 Duplicate Elimination, pages 17-19, and particularly the disclosure at the top of page 18 that the formal definition for a similar collection includes requirements that that the pages are similar, and also that the links are similar, meaning that each of the pages should have at least one parent in their corresponding collection that are also similar pages; see also disclosure of the use of hyperlinks in categorizing documents, section 4.2.3 Enhanced Categorization Using Hyperlinks, specifically Radius-Two Specialization, which uses co-citation as a criterion for classification, on page 24).

It would have been obvious to one of ordinary skill in the art at the time of the invention to define templates with regard to similar pagelets and also similar links, since this would help to ensure that the templates being scrutinized are in fact similar, given the fact that one of the major difficulties in detecting replicated collections is that many replicas may not be strictly identical to each other (page 17).

17. Regarding claim 11, **Broder et al.** teaches an apparatus as claimed, comprising:

a) a user interface (see disclosure that the system can be used for filtering the results of Web searches, said web searches requiring a user interface, Abstract);

b) a user interface/event manager communicatively coupled to the user interface (see disclosure that the system can be used for filtering the results of Web searches, said web searches requiring an event handler to respond to user requests, Abstract);

c) a generic data gathering device (see disclosure that the system can be applied to a group of documents found by the AltaVista spider, section 1 Introduction, beginning on page 2);

d) a generic information retrieval application, communicatively coupled to the user interface/event manager (see disclosure that the system can be applied to a group of documents found by the AltaVista spider, section 1 Introduction, beginning on page 2); and

e) a data cleaning application communicatively coupled to the generic data gathering application and to the generic information retrieval application for

i) decomposing each page of a set of text documents into one or more pagelets (see disclosure that documents are decomposed into shingles, analogous to the claimed pagelets, in section 2 Defining Similarity of Documents, beginning on page 3);

ii) identifying all pagelets belonging to templates (see disclosure that common shingles were nearly all mechanically generated, including shared header or footer information on a large number of automatically generated pages, i.e. forms,

analogous to the claimed templates, in section 5.1 Common Shingles, beginning on page 7); and

iii) eliminating the template pagelets from a data set (see section 5.1 Common Shingles, beginning on page 7, particularly the disclosure on page 8 that common shingles either have no effect on the overall resemblance of the documents or they have the effect of creating a false resemblance between two basically dissimilar documents, and so common shingles are ignored).

Broder et al. does not explicitly teach a method wherein a template comprises a collection of pagelets satisfying the requirements that all pagelets are identical or almost identical, and every two pages owning pagelets are reachable one from the other via other pages also owning pagelets in the collection.

Huang, however, teaches a method for eliminating duplicate web pages wherein a template comprises a collection of pagelets satisfying the requirements that all pagelets are identical or almost identical, and every two pages owning pagelets are reachable one from the other via other pages also owning pagelets in the collection (see section 3.5.2 Duplicate Elimination, pages 17-19, and particularly the disclosure at the top of page 18 that the formal definition for a similar collection includes requirements that that the pages are similar, and also that the links are similar, meaning that each of the pages should have at least one parent in their corresponding collection that are also similar pages; see also disclosure of the use of hyperlinks in categorizing documents, section 4.2.3 Enhanced Categorization Using Hyperlinks, specifically Radius-Two Specialization, which uses co-citation as a criterion for classification, on page 24).

It would have been obvious to one of ordinary skill in the art at the time of the invention to define templates with regard to similar pagelets and also similar links, since this would help to ensure that the templates being scrutinized are in fact similar, given the fact that one of the major difficulties in detecting replicated collections is that many replicas may not be strictly identical to each other (page 17).

18. Regarding claim 13, **Broder et al.** teaches a computer readable medium including computer instructions for driving a user interface as claimed, the computer instructions comprising instructions for cleaning, by operations of a computer system, a set of text documents to minimize violations of a predetermined set of Hypertext Information Retrieval rules (see section 1 Introduction, beginning on page 2; see also section 5.1 Common Shingles, beginning on page 7, particularly the disclosure on page 8 that common shingles either have no effect on the overall resemblance of the documents or they have the effect of creating a false resemblance between two basically dissimilar documents, and so common shingles are ignored), wherein the cleaning step comprises the steps of:

a) decomposing each page of the set of text documents into one or more pagelets (see disclosure that documents are decomposed into shingles, analogous to the claimed pagelets, in section 2 Defining Similarity of Documents, beginning on page 3);

b) identifying all pagelets belonging to templates (see disclosure that common shingles were nearly all mechanically generated, including shared header or footer information on a

large number of automatically generated pages, i.e. forms, analogous to the claimed templates, in section 5.1 Common Shingles, beginning on page 7); and

c) eliminating the template pagelets from a data set (see section 5.1 Common Shingles, beginning on page 7, particularly the disclosure on page 8 that common shingles either have no effect on the overall resemblance of the documents or they have the effect of creating a false resemblance between two basically dissimilar documents, and so common shingles are ignored).

Broder et al. does not explicitly teach a method wherein a template comprises a collection of pagelets satisfying the requirements that all pagelets are identical or almost identical, and every two pages owning pagelets are reachable one from the other via other pages also owning pagelets in the collection.

Huang, however, teaches a method for eliminating duplicate web pages wherein a template comprises a collection of pagelets satisfying the requirements that all pagelets are identical or almost identical, and every two pages owning pagelets are reachable one from the other via other pages also owning pagelets in the collection (see section 3.5.2 Duplicate Elimination, pages 17-19, and particularly the disclosure at the top of page 18 that the formal definition for a similar collection includes requirements that that the pages are similar, and also that the links are similar, meaning that each of the pages should have at least one parent in their corresponding collection that are also similar pages; see also disclosure of the use of hyperlinks in categorizing documents, section 4.2.3 Enhanced Categorization Using Hyperlinks, specifically Radius-Two Specialization, which uses co-citation as a criterion for classification, on page 24).

It would have been obvious to one of ordinary skill in the art at the time of the invention to define templates with regard to similar pagelets and also similar links, since this would help to ensure that the templates being scrutinized are in fact similar, given the fact that one of the major difficulties in detecting replicated collections is that many replicas may not be strictly identical to each other (page 17).

19. Regarding claim 20, **Broder et al.** teaches a computer readable medium including computer instructions for driving a user interface as claimed, the computer instructions comprising instructions for cleaning, by operations of a computer system, a set of text documents to minimize violations of a predetermined set of Hypertext Information Retrieval rules (see section 1 Introduction, beginning on page 2; see also section 5.1 Common Shingles, beginning on page 7, particularly the disclosure on page 8 that common shingles either have no effect on the overall resemblance of the documents or they have the effect of creating a false resemblance between two basically dissimilar documents, and so common shingles are ignored), wherein the cleaning step comprises the steps of:

a) decomposing each page of the set of text documents into one or more pagelets (see disclosure that documents are decomposed into shingles, analogous to the claimed pagelets, in section 2 Defining Similarity of Documents, beginning on page 3);

b) identifying all pagelets belonging to templates (see disclosure that common shingles were nearly all mechanically generated, including shared header or footer information on a

large number of automatically generated pages, i.e. forms, analogous to the claimed templates, in section 5.1 Common Shingles, beginning on page 7); and

c) eliminating the template pagelets from a data set (see section 5.1 Common Shingles, beginning on page 7, particularly the disclosure on page 8 that common shingles either have no effect on the overall resemblance of the documents or they have the effect of creating a false resemblance between two basically dissimilar documents, and so common shingles are ignored), wherein the identifying step comprises:

i) calculating a shingle value for each page and for each pagelet in the document set (see disclosure that documents are decomposed into shingles, analogous to the claimed pagelets, in section 2 Defining Similarity of Documents, beginning on page 3; see also disclosure of the fingerprint function, refered to as the shingle value, page 6); and

ii) sorting the pagelets by their shingle values into clusters (see disclosure in section 4 Algorithms, whereby similar documents are clustered, page 6).

Broder et al. does not explicitly teach a method wherein a template comprises a collection of pagelets satisfying the requirements that all pagelets are identical or almost identical, and every two pages owning pagelets are reachable one from the other via other pages also owning pagelets in the collection.

Huang, however, teaches a method for eliminating duplicate web pages wherein a template comprises a collection of pagelets satisfying the requirements that all pagelets are identical or almost identical, and every two pages owning pagelets are reachable one from the other via other pages also

owning pagelets in the collection (see section 3.5.2 Duplicate Elimination, pages 17-19, and particularly the disclosure at the top of page 18 that the formal definition for a similar collection includes requirements that that the pages are similar, and also that the links are similar, meaning that each of the pages should have at least one parent in their corresponding collection that are also similar pages; see also disclosure of the use of hyperlinks in categorizing documents, section 4.2.3 Enhanced Categorization Using Hyperlinks, specifically Radius-Two Specialization, which uses co-citation as a criterion for classification, on page 24).

It would have been obvious to one of ordinary skill in the art at the time of the invention to define templates with regard to similar pagelets and also similar links, since this would help to ensure that the templates being scrutinized are in fact similar, given the fact that one of the major difficulties in detecting replicated collections is that many replicas may not be strictly identical to each other (page 17).

20. Regarding claims 2 and 14, **Broder et al.** additionally teaches a method and computer readable medium as claimed, wherein the set of text documents comprises a collection of HTML pages (see disclosure in the Abstract that the disclosed invention is applied to every document on the World Wide Web, page 1).

21. Regarding claims 7 and 19, **Broder et al.** additionally teaches a method and computer readable medium as claimed, wherein the step of identifying all pagelets belonging to templates comprises the steps of:

a) calculating a shingle value for each page and for each pagelet in the set of documents (see disclosure of the calculation of fingerprint values on the shingles, in section 3 Estimating the Resemblance and the Containment, beginning on page 4);

b) eliminating identical pagelets belonging to duplicate pages (see section 5.1 Common Shingles, beginning on page 7, particularly the disclosure on page 8 that common shingles either have no effect on the overall resemblance of the documents or they have the effect of creating a false resemblance between two basically dissimilar documents, and so common shingles are ignored);

c) sorting the pagelets by their shingle value into clusters (see disclosure of the clustering procedure in section 4.1 The Clustering Algorithm, on page 7);

d) enumerating the clusters (see disclosure of the clustering procedure in section 4.1 The Clustering Algorithm, on page 7); and

e) outputting a representation corresponding to the pagelets belonging to each cluster (see disclosure of the clustering procedure in section 4.1 The Clustering Algorithm, on page 7).

22. Claims 4 and 16 are rejected under 35 U.S.C. 103(a) as being unpatentable over **Broder et al.** ("Syntactic Clustering of the Web") in view of **Huang** ("A Survey on Web Information Retrieval Technologies") as applied to claims 1, 2, 7-9, 11, 13, 14, 19 and 20 above, and further in view of **Chakrabarti et al.** ("Enhanced Topic Distillation using Text, Markup Tags and Hyperlinks").

23. Regarding claims 4 and 16, **Broder et al.** and **Huang** teach a system and apparatus substantially as claimed.

Neither **Broder et al.** nor **Huang** explicitly teaches a system and apparatus wherein the decomposing step comprises the claimed steps.

Chakrabarti et al., however, teaches a system and apparatus wherein the decomposing step comprises the steps of:

a) parsing each text document into a parse tree that comprises at least one node (see disclosure that each HTML page is a Document Object Model (DOM) tree, p. 210, under section 3 Proposed Model and Algorithms);

b) traversing the at least one node of the tree (see disclosure that each HTML page is a Document Object Model (DOM) tree, p. 210, under section 3 Proposed Model and Algorithms; see also Figure 4, page 211, illustrating the finished tree wherein pagelets have been pushed to the leaves of the tree; see also section 3.2 Segmentation and Smoothing, page 211);

c) determining if one of the at least one node comprises a pagelet (see disclosure that each HTML page is a Document Object Model (DOM) tree, p. 210, under section 3 Proposed Model and Algorithms; see also Figure 4, page 211, illustrating the finished tree wherein pagelets have been pushed to the leaves of the tree; see also section 3.2 Segmentation and Smoothing, page 211); and

d) outputting a representation corresponding to the one of the at least one node if it comprises a pagelet (see disclosure that each HTML page is a Document Object

Model (DOM) tree, p. 210, under section 3 Proposed Model and Algorithms; see also Figure 4, page 211, illustrating the finished tree wherein pagelets have been pushed to the leaves of the tree; see also section 3.2 Segmentation and Smoothing, page 211).

It would have been obvious to one of ordinary skill in the art at the time of the invention to decompose an HTML document to arrive at a list of pagelets through the use of a parse tree, since it is important to bring in additional sources of information (like a tag tree structure) where possible, to combat topic drift and clique attacks (see page 210, col. 1, last paragraph).

24. Claims 10 and 12 are rejected under 35 U.S.C. 103(a) as being unpatentable over **Broder et al.** ("Syntactic Clustering of the Web") in view of **Huang** ("A Survey on Web Information Retrieval Technologies") as applied to claims 1, 2, 7-9, 11, 13, 14, 19 and 20 above, and further in view of **Rodeheffer et al.** (U.S. Patent 6,614,764).

25. Regarding claims 10 and 12, **Broder et al.** and **Huang** teach a system and apparatus substantially as claimed, further comprising:

a) a pagelet identifier, communicatively coupled to the data cleaning application (see **Broder et al.** disclosure that documents are decomposed into shingles, analogous to the claimed pagelets, in section 2 Defining Similarity of Documents, beginning on page 3);

b) a hypertext parser, communicatively coupled to the pagelet identifier (see **Broder et al.** disclosure in the Abstract that the disclosed invention is applied to every document on the World Wide Web, page 1; see also disclosure that documents are decomposed into shingles, in section 2 Defining Similarity of Documents, beginning on page 3);

c) a template identifier, communicatively coupled to the data cleaning application (see **Broder et al.** disclosure that common shingles were nearly all mechanically generated, including shared header or footer information on a large number of automatically generated pages, i.e. forms, analogous to the claimed templates, in section 5.1 Common Shingles, beginning on page 7); and

d) a shingle calculator, communicatively coupled to the data cleaning application (see **Broder et al.** disclosure of shingle construction, section 2 Defining Similarity of Documents, beginning on page 3).

Neither **Broder et al.** nor **Huang** explicitly teaches a system and apparatus comprising a Breadth First Search (BFS) algorithm.

Rodeheffer et al., however, teaches the Breadth First Search (BFS) teachnique (see col. 34, lines 43-62).

It would have been obvious to one of ordinary skill in the art at the time of the invention to incorporate a Breadth First Search (BFS) algorithm, since this would produce a spanning tree in which the path from each node to the root is as short as possible, and generally, shorter paths are better. Furthermore, the breadth-first search is also efficient (see col. 34, lines 51-62).

Response to Arguments

26. Applicant's arguments filed 24 November 2004 have been fully considered but they are not persuasive.

27. In response to the Applicants' arguments regarding the pending claim rejections under 35 U.S.C. § 112, the examiner has found the arguments persuasive, although as discussed above with regard to the drawings, the change in interpretation of the claims necessitates a correction to Figure 6.

28. In response to the Applicants' arguments regarding the claim rejections of independent claims 1, 9, 11 and 13, the examiner notes the amendment to these claims. However, these claims, including the newly added limitations, have been rejected herein based in part upon newly discovered prior art.

Furthermore, the newly amended limitations raise an issue regarding the definiteness of the claims. See discussion above under 35 U.S.C. § 112.

29. In addition, said newly discovered prior art also reads upon claims 8 and 20, these claims having previously been indicated by the examiner as containing allowable subject matter. This indication of allowable subject matter has been withdrawn, and a rejection based in part upon this newly discovered prior art is included herein. For this reason, this Office action is non-final.

Conclusion

30. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure.

Shivakumar et al. ("SCAM: A Copy Detection Mechanism for Digital Documents") teaches a new scheme for detecting copies of documents based upon comparison of the word frequency occurrences of a new document against those of registered documents.

Bharat et al. ("Mirror, Mirror on the Web: A Study of Host Pairs with Replicated Content") teaches a technique for detecting mirroring hosts from large sets of URLs.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Luke S. Wassum whose telephone number is 571-272-4119. The examiner can normally be reached on Monday-Friday 8:30-5:30, alternate Fridays off.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, John E. Breene can be reached on 571-272-4107. The fax phone number for the organization where this application or proceeding is assigned is 703-872-9306.

In addition, INFORMAL or DRAFT communications may be faxed directly to the examiner at 571-273-4119.

Customer Service for Tech Center 2100 can be reached during regular business hours at (571) 272-2100, or fax (703) 872-9306.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

Luke S. Wassum
Primary Examiner
Art Unit 2167

lsw
24 February 2005